doi：10.12452/j.fxcsxb.26012901

AI驱动下中药化学成分的智能分析与化学空间拓展

卢志鹏，董莹莹，杜　柯，单进军，谢　彤*

（南京中医药大学儿科研究所，儿童健康与中医药省高校重点实验室，江苏　南京　210023）

摘　要：中药复杂体系包含小分子、蛋白质、多糖、高阶自组装体等多种化学成分。当前，这些成分的分析受限于谱库覆盖度不足与人工解析效率低下，制约了物质基础研究的深度。人工智能（AI）凭借强大的多源数据整合与深度表征学习能力，推动中药化学成分解析从“经验驱动”向“数据驱动”的新范式转变。该文系统综述了AI在中药成分分析中的学习范式和核心架构，重点剖析了分子表征学习方法、谱图数据处理策略以及在智能分析与化学空间拓展方面的前沿应用。最后，探讨了数据标准化、数据共享和多模态整合等关键挑战与未来发展路径。

关键词：人工智能；中药化学成分；深度学习；表征学习；智能分析

中图分类号：O657；R284　　**文献标识码：**A　　**文章编号：**1004-4957(XXXX)XX-0001-14

AI-driven Intelligent Characterization and Chemical Space Expansion for Traditional Chinese Medicine

LU Zhi-peng，DONG Ying-ying，DU Ke，SHAN Jin-jun，XIE Tong*

（Institute of Pediatrics，Children's Health and Traditional Chinese Medicine Provincial Key Laboratory，Nanjing University of Chinese Medicine，Nanjing 210023，China）

Abstract：The compositional analysis of complex material systems in traditional Chinese medicine（TCM）—including small molecules，proteins，polysaccharides，and higher-order self-assembled structures—has long been constrained by insufficient spectral library coverage and the bottleneck of manual interpretation，thereby limiting the standardization and depth of research into their material basis. Leveraging its strong capability for multi-source data integration and deep representation learning，artificial intelligence（AI）is driving a paradigm shift in TCM constituent analysis from experience-driven approaches to data-driven methodologies. This review systematically surveys learning paradigms，core architectures，and representative tasks of AI-based TCM constituent analysis，with a particular focus on molecular representation learning methods，spectroscopic data processing strategies，and recent advances in complex multidimensional component identification and chemical space expansion. Finally，key challenges and future directions are discussed，including data standardization，multimodal integration，and model interpretability. From an interdisciplinary perspective，this review aims to provide methodological support for the modernization and high-quality development of traditional Chinese medicine.

Key words：artificial intelligence；traditional Chinese medicine（TCM）；deep learning；representation learning；component characterization

人工智能（AI）通过机器学习赋能中医药领域，在智能诊疗、组方优化、临床前智能研发、靶点发现与验证等方面发挥重要作用。然而，针对中药化学成分智能分析的AI系统研发仍处于起步阶段。为了推进其发展，本文聚焦于以下关键问题展开讨论：（1）深度学习的架构；（2）分子表征的技术；（3）谱图表征的技术；（4）AI在中药化学成分智能分析中的具体应用；（5）研发通用智能系统的挑战和未来展望。本文所阐述的创新技术可作为系列协同工具，为构建中药智能分析系统提供参考。

收稿日期：2026-01-29；**修回日期：**2026-04-17

基金项目：江苏省自然科学基金面上项目（BK20241915）；江苏省中医药科技计划项目（MS2022002）

*通讯作者：谢　彤，博士，教授，研究方向：中医药代谢组学、药效物质发现的新技术与新方法，E-mail：xietong@njucm.edu.cn

1 机器学习的架构及应用场景

为了方便阐述，首先厘清机器学习中算法、架构、模型、软件框架和智能体/智能系统的概念（表1）。文中的“架构”指为解决特定问题而设计的、包含数据流程与算法组合的计算框架，区别于PyTorch和TensorFlow等实现框架的软件框架。架构的选择至关重要，其有效性取决于待解决问题的性质和数据特征。

表1 算法、架构、模型、软件框架和智能体/智能系统的概念

Table 1 Concept of algorithms, models, frameworks and agent/ intelligent systems

Concept	Description	Example
Algorithm（算法）	配置深度学习模型的规则或指令。是一种基于数学层面的计算方法	反向传播算法、随机梯度下降法、Adam等
Architecture（架构）	侧重于模型的整体结构设计。指针对某一类任务的通用解决方案，在该方案中，特定模块可以采用不同算法。	生成对抗网络（GAN）、Seq2Seq、Transformer等
Model（模型）	指基于特定网络架构，通过数据训练得到的参数化实例，可直接用于具体任务	Alphafold2、ChatGPT、DeepSeek等
Framework（框架）	支撑模型构建、训练与部署的软件系统，提供算子库、自动求导和运行管理等功能	PyTorch、TensorFlow、JAX等
Agent/Intelligent System（智能体/智能系统）	指能够感知环境、进行决策并执行行动以实现目标的集成系统，通常包含模型、规划器、工具调用等组件	ToolUniverse、Devin、AutoGPT等

1.1 机器学习范式/策略

传统的学习范式包括监督学习和无监督学习。监督学习使用大量有标签的数据训练模型，但数据标注成本很高。为此，发展出半监督学习，同时利用少量有标签数据与大量未标注数据来增强模型训练。无监督学习则使用未标注的数据训练模型[1]。然而其学习效果难以评价。由此发展出自监督学习，通过随机掩码等方式，在无标签数据上建立监督信号进行学习。对比学习是自监督学习的实现手段之一，通过构造正负样本对来学习。强化学习（RL）则通过奖励或惩罚函数学习最优决策，广泛用于化学空间探索等方向[2-3]。迁移学习是一种跨任务的学习策略，即将某领域上学习到的知识，应用到另一个相关领域，其能较好地解决中药标注数据稀缺的问题，如预测中药药效靶点、识别中药材图像[4]等。表2汇总了各机器学习方法的特点及代表性模型。需要指出的是，上述范式在实际应用中并非孤立应用，许多模型采用混合架构，结合多种学习策略来解决问题。

表2 机器学习范式/策略对比

Table 2 Comparison of machine learning paradigms and strategies

Learning paradigm/Strategy	Data type and representative model
监督学习	大量有标签的数据进行学习，如SIRIUS[5]
无监督学习	无标签数据进行学习，如Spec2Vec[1]
半监督学习	少量标签+大量无标签数据进行学习，如FixMatch[6]
自监督学习	海量无标签的数据进行学习，如DreaMS[7]、MolLM[8]
强化学习	通过奖励/惩罚信号来确定最优决策，如REINVENT[9]
迁移学习	源领域学习，迁移至目标域，如ChemBERTa[10]

1.2 机器学习方法

1.2.1 按技术分：传统机器学习和深度学习

1.2.1.1 *传统机器学习* 传统机器学习是依赖于结构化数据和人工定义特征的经典方法。例如用红外光谱鉴别中药材产地时，选取特定波数的吸光度或峰强度作为特征，构建判别分析模型[11]。然而，传统机器学习泛化能力有限，即便引入正则化等技术，在处理复杂的非线性数据时仍显不足。为此，研究者转向将特征提取与模型训练融为一体的端到端深度学习模型。深度学习（DL）通过构建多层神经网络，能够自动从原始数据中学习特征，避免了人工特征工程的局限。

1.2.1.2 *深度学习* 常见的DL架构包括前馈神经网络（FNN）、卷积神经网络（CNN）、循环神经网络（RNN）。FNN是基础架构，在其中引入卷积核和池化机制即形成CNN。CNN在图像任务中表现很好，如DeepeR利用CNN处理拉曼光谱成像[12]。RNN专用于处理文本、时间等序列数据。为解决RNN梯度消失问题，发展出长短时记忆网络（LSTM）和门控循环单元（GRU）。LSTM可捕捉序列中的长期

依赖关系，例如Srisongkram等[13]利用双向LSTM从SMILES序列中提取语义信息，提升模型预测准确率。GRU是LSTM的简化变体，在保持性能的同时具有较高计算效率。如Sugumaran等[14]基于GRU预测药物-靶点相互作用。除上述架构外，图神经网络（GNN）和基于自注意力机制的Transformer架构，近年来在化学与中药领域展现出强大潜力。

GNN是处理图结构的神经网络，大多数基于消息传递机制构建。由于化学分子即可以被表示为图结构，GNN被广泛应用于分子性质预测和分子生成等任务中[15]。Transformer最初用于序列数据处理，而后衍生多种子架构。如处理图结构的图Transformer、处理图像的Vision Transformer架构以及以BERT为代表的预训练语言模型，并在此基础上进一步发展出大语言模型（LLMs）。例如Masood等[16]将预训练BERT嵌入学习框架，提升化合物毒性预测能力。LLMs在处理自然语言时展现出强大能力，如运用LLMs挖掘中医药文献与古籍中的知识[17]。基于Transformer的 LLMs亦可用于新分子生成任务。

1.2.2　按照模型目的分类：判别模型和生成模型

1.2.2.1　*判别模型*　判别模型通过已知样本特征建立分类规则，实现对新样本的类别预测，在中药成分分析中应用广泛。例如Zhou等[18]通过机器学习模型，对Panax quinquefolius L. 的品种与产地进行有效区分；高美美等[19]则基于HPLC指纹图谱及化学模式识别方法对不同产地杉木叶进行质量分析与分类。

1.2.2.2　*生成模型*　相较于判别模型，生成模型具备创造能力，是当前AI领域的热点方向。在中药成分分析中，MSNovelist[20]和MSGo[21]等工具已将生成模型用于从质谱谱图直接推导全新分子骨架：前者基于RNN将分子指纹解码生成SMILES序列，后者通过GNN生成分子图。从更系统的角度来看，当前主流的架构包括生成对抗网络（GAN）[22]、变分自编码器（VAE）[23]、标准化流模型（NF）[24]和扩散模型（DM）[25]。

研究者还不断探索跨模态的集成架构，通过桥接不同类型的数据增强生成能力。例如，MolEdit整合了不同分子类型的数据，支持多样化分子的生成[26]；DiffNovo和DiffMS将质谱谱图和分子结构映射到统一框架中，通过谱图直接生成分子结构[27]。需要指出的是，在“谱图推测结构”的任务中，判别模型依赖已知数据库进行匹配检索预测结构，而生成模型可从零创造全新分子，不受数据库限制，对中药未知成分探索具有重要意义。

1.3　可解释人工智能

随着AI模型日益复杂，庞大的参数规模导致模型难以理解，陷入“黑箱”困境。因此，可解释人工智能（XAI）成为重要研究方向，其目的是使决策过程变得可被理解[28]。目前，针对中药领域的集成XAI模型不多见。但是在实际运用端，通过图解释性技术（如Hierarchical Grad-CAM）已能有效揭示影响模型的关键因素。相关的中药AI模型也集成了可解释性的增强工具，帮助解释模型的输出，如TCMNPS具备网络药理学的可视化功能。

2　分子表征的方法

分子表征指将分子的结构和属性等信息转化为被计算机处理的数据形式。依赖领域内的已知知识，以人工介入方式来确定分子表征的方法称之为人工特征。通过模型让机器自动从原始数据（如SMILES）中学习分子特征的过程称为表征学习。需要说明的是，嵌入和表征学习在许多论文中被交替使用，但在具体的学术语境中，模型输出的向量通常被称为嵌入，而训练模型的过程被称为表征学习。

2.1　传统的分子表征方法

线形符号（如SMILES、SELFIES）可直接被RNN、Transformer等模型读取，但存在非唯一性及表征不稳定的问题。分子指纹将复杂拓扑结构映射为定长向量，提升了化学空间检索的效率，但难以反还为原始分子结构。分子描述符也是常用的分子表征方法。RDKit和PaDEL等工具可批量计算数百种分子描述符；对于三维结构，可通过小波变换不变量、固体谐波小波散射变换等非参数数学方法，获得平移和旋转不变的特征向量作为3D描述符。然而，这些人工特征在面对复杂结构时易丢失信息，泛化能力有限，因此许多研究转向表征学习来处理化学结构。

2.2 基于表征学习的方法

表征学习通过不同维度对分子进行数学建模以学习分子特征，主要分成3个方向：语言学习模型学习（针对分子的一维序列）、GNN和图Transformer学习（针对分子的二维拓扑图）、几何深度学习（针对分子的三维空间坐标）。

2.2.1 基于序列的表征学习 化学结构与自然语言（NLP）在语义逻辑上具有相似性[29]。借鉴NLP处理方式，将各类分子转化为序列表示，即Tokens表征：小分子化合物编码为SMILES等线性符号；多肽、蛋白质、多糖等，映射为氨基酸序列或糖基序列。目前，主流的序列架构有RNN、Transformer和BERT。在小分子领域，代表性的语言学习模型有ChemBERTa[10]和Mol2vec[30]等。蛋白质大语言模型发展尤为迅速，涌现出如AlphaFold2和ESMFold等代表性模型；针对复杂的翻译后修饰，研究者引入额外的特征嵌入以增强学习能力（如DeepGlycanSite和DeepGlyco等[31]）。相比之下，多糖因非线性分支特征，其序列建模的难度较大。针对这一挑战，GlycoBERT和GlycoBART通过引入树状编码机制，实现了复杂糖链的深度拓扑表征[32]。然而，中药复杂组分中存在高度重复和高级折叠的巨型多糖，建模难度大，需要进一步结合粗粒化模型探索改进策略。

2.2.2 基于图结构的表征学习 分子结构可视为图进行表征学习：原子作为节点，化学键作为边。对于结构复杂的大分子，则可简化至残基水平，将单糖或氨基酸残基视为节点，糖苷键或肽键视为边。图Transformer是处理此类结构的主流架构。

小分子的表征学习是图Transformer应用最为广泛的领域，研究者开发出各种改进策略。例如DrugEx v3设计了基于邻接矩阵的位置编码，提高了图Transformer对分子图的理解[33]；MolE利用原子标识符和拓扑距离矩阵作为输入，为模型提供了更详细的结构信息[34]；KPGT采用知识引导策略，使模型能够捕捉全局信息和远程节点关系[35]；Massformer则将功能基团信息与图结构结合以充分表征分子[36]。

多糖的生物学功能由其单糖组成、顺序、连接方式和分支结构决定，这些信息包含于二维拓扑图中。因此，图结构是研究多糖的有效方法，主要模型有GNNGLY[37]、SweetNet[38]和GIFFLAR[39]等。GNNGLY在原子层面对糖链进行表征学习[37]；SweetNet将单糖分子和糖苷键均视为节点，学习多糖的非线性分支[38]；GIFFLAR通过“组合复合体”概念统一表征原子、键和单糖多个层次，并借助高阶拓扑聚合捕捉复杂空间结构特征[39]。蛋白质的图结构表征，主要集中在三维结构而非二维拓扑图，因此在“2.2.3”中进行讨论。

2.2.3 三维结构的几何深度学习 几何深度学习（GDL）是处理分子3D信息的主要方法。它从原子坐标提取特征，在保证原子排列不变性的同时，满足欧几里得变换群E（3）（旋转、平移和反射）的等变性或不变性。

对于构象已知的分子，GDL可用于分子表征，从而有效预测其理化性质与生物功能。GDL的主流架构为几何图神经网络（G-GNNs），该网络通过不变层或等变层处理空间坐标信息。在小分子领域，不变架构包括DimeNet[40]和GemNet[41]等，用于预测分子能量、结合亲和力、热力学性质等；等变架构包括张量场网络TFN[42]、PaiNN[43]和NequIP[44]等，用于预测偶极矩、振动模式和原子力等任务。在蛋白质领域，DeepRank-GNN[45]和MaSIF[46]等不变性模型，用于亲和力结合强度预测等；EquiPPIS[47]和ESMFold等等变架构用于预测相互作用、蛋白功能位点等。

若构象未知，GDL则演变成为生成模型。对于小分子，主要任务是从二维拓扑生成合理的3D构象，典型模型包括ConfGF[48]和GCDM[49]。对于大分子（主要是蛋白质），主要从序列信息生成空间结构，代表性模型如AlphaFold和ESMFold。然而，由于多糖具有高柔性、复杂支链和多样化学修饰的特点，GDL在多糖及糖肽领域的应用仍面临挑战。因此，可结合动态采样等方法构建糖类GDL模型，以实现其空间精准表征。

2.2.4 分子表征的多模态整合 将不同维度数据整合到统一的学习框架中，通过信息互补实现对分子的全面理解，已成为当前研究大的重要方向。在蛋白质领域，通常整合一维序列特征和三维空间特征，以增强功能与属性预测，如AlphaFold等。在小分子领域，则注重整合拓扑结构、三维构象和文本描述等多模态信息。例如MolLM通过整合分子图和序列进行自监督学习，从而增强分子性质的预测能

力[8]。此外，有研究将物理约束和几何先验信息融入表征学习中以预测高阶结构，这对跨尺度的结构研究尤为重要[50]。

此外，还可整合不同类型的分子。如FuncBind框架利用神经场构建了覆盖小分子、大环肽和抗体CDR环的统一模型[51]。T5ProtChem构建了跨生物大分子和小分子的统一语言模型[52]。此类模型显著增强了对跨尺度化学信息的处理能力。

3　谱图表征及处理方法

谱图表征是通过DL将谱图物理信号转化为计算机可理解的特征向量。目前，谱图表征的策略主要有两种：一是预训练-微调策略，即利用MIST[53]和MS2DeepScore[54]等已预训练的成熟工具，通过迁移学习来适配下游任务；二是从头训练，即针对特定需求，通过全参数训练构建专用模型（图1）。

中药成分分析通常涉及质谱（含碰撞截面积CCS）、光谱以及色谱等信息。其中，CCS与色谱数据的输入特征相对简单，故本文不作赘述。下文将立足于质谱与光谱技术，重点阐述其表征学习的新进展与应用。

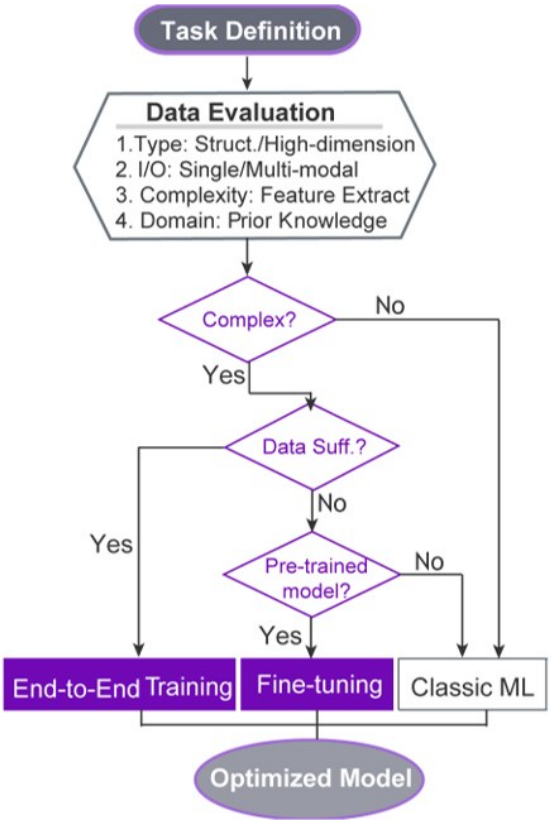

图1　谱图表征学习模型开发的流程决策图

Fig. 1　Decision workflow for representation learning model

3.1　质谱数据的表征与预处理策略

3.1.1　谱图表征和编码方法　质谱数据的表征常采用分箱、词嵌入、正弦波质荷比嵌入的方法。分箱操作将质量范围划分成若干区间（Bins），产生的向量作为DL的输入，如MS2DeepScore[54]和MS2Query[55]即采用该方法。其中，分箱的箱宽设置很重要，箱宽过大会丢失信息，太窄会增加计算量。为解决此问题，AHLF采用双向量谱图输入策略，在降低数据维度的同时，保留了关键信息[56]。词嵌入借鉴了NLP思想，将碎片离子和中性丢失转化为离散词语生成谱图特征向量，如Spec2Vec[1]。正弦波质荷比嵌入，利用正弦波函数编码谱图，保留了高分辨质谱信息，是当前研究的主流，如CSU-MS2和SpecEmbedding[57-58]等均使用该方法。此外，还包括多频率傅里叶编码[7]及融合领域内知识等增强策略[53]。

针对质谱成像数据，需将原始谱图数据和空间信息转化为低维嵌入向量。例如，msiPL模型[59]通

过VAE以无监督方法获得谱图嵌入，MassNet[60]在其基础上叠加分类器实现自动分类，DeepION[61]利用对比学习自监督生成离子图像表征。此外，有研究引入物理约束进行多模态融合，以整合ToF-SIMS与MALDI数据，提升了化学表征的效率[62]。

3.1.2 数据获取途径 质谱数据主要来源于实验采集、公共数据库和计算机模拟。公共数据库包括MassBank、GNPS、MoNA 、mzCloud和NIST等小分子库，以及PRIDE、PeptideAtlas、NIST Protein Database等蛋白质库。然而，现有数据库缺乏中药成分的谱图数据，因此在构建模型时，除了实验数据外，还可利用CFM-ID[63]、SIRIUS[5]、Prosit[64]等预测工具，针对虚拟候选化合物库生成模拟质谱图来扩充数据。

3.1.3 数据清洗和数据增强策略 为了消除不同来源数据的差异，通过需要对质谱数据进行清洗，处理工具包括OpenMS 3[65]、NeatMS[66]等。数据清洗后，还会通过数据增强策略以提高模型的鲁棒性，即对原始数据进行人为变换以增加训练样本多样性，常见的数据增强方式包括对谱图随机注入噪声、平移（改变m/z值）或缩放谱图（调整强度分布），来模拟不同实验条件对数据的影响，从而使模型在面对不同仪器偏差、实验噪声时依然能识别出样本。

3.2 光谱数据表征与预处理策略

3.2.1 谱图表征和编码方法 光谱数据的表征方法经历了从人工特征工程到自动化端到端学习的转变。人工特征工程将原始谱图的化学位移或波数作为输入，例如ECDFormer将光谱图的峰位、峰强和半峰宽作为输入进行模型训练[67]。但该方法依赖领域内知识，易丢失隐藏的信息。

其次是张量输入法，如CReSS和SpecGNN将光谱信号转化为一维数组输入模型，实现端到端的学习；对于二维光谱，可将谱图拆分成两个一维数组输入模型[68]。采用张量法编码谱图时，采样点的平衡至关重要。采样点过多会产生冗余值，采样点过少又会丢失关键信息。针对此问题，研究者利用自注意力机制分块输入的方法，在保持信息量的同时显著降低计算负担[69]。此外，为了解决谱图统一格式的问题，研究者利用线性插值算法将不同分辨率的光谱图重排到统一坐标上[70]。在NMR领域，利用AI解决谱图重建取得了显著成功。如JTF-Net和FID-Net等直接将时域和频域信号作为高维复数张量输入训练模型[71]，进而有效学习谱图完整信息。

再者，有研究将光谱图视为图像进行处理，如将NMR谱图转化成图像输入CNN[72]，通过马尔可夫转移场处理拉曼光谱，或将多模态光谱拼接并重塑成二维图像[73]。但是，此类基于图像表征的方法主要集中在分类任务。

3.2.2 数据获取途径 与质谱类似，标注的光谱数据面临稀缺挑战。除了实验与数据库外，可基于谱图生成获得模拟谱图。如在NMR中，通过Voigt线型物理模型，随机生成一维谱图[68]；也可基于物理规则建立谱图片段生成器，以此扩充训练数据规模[74]。

3.2.3 数据清洗和数据增强策略 为了排除光谱图中的干扰，通过预处理工具进行数据清洗。在NMR领域，ARTINA[75]和DEEP Picker[68]模型能自动完成噪音识别和峰解卷积。对于拉曼光谱，ResNet或U-Net可实现全自动基线校正和去噪[76]。数据清洗后，采用数据增强策略提升模型泛化能力，如在红外光谱区添加自适应噪声、改变水平位移、进行高斯平滑[69]；在拉曼光谱中模拟引入暗噪声、基线漂移等干扰[77]。图2对学习范式、模型架构及实际应用流程进行了系统整合。

4 构建中药成分智能分析与新物质发现的通用型AI平台

现有综述已经系统总结了AI在中药质量评价、作用机制解析以及精准治疗等方面的应用进展。然而，在中药成分分析领域，如何充分利用先进AI技术实现中药成分的快速、准确表征及发现未知新物质仍然是一个挑战。

4.1 谱图的智能解析（谱图→结构）

传统的谱图解析依赖专家经验和有限的谱库，制约了中药复杂组分的解析。AI能自动学习谱图特征，进而自动化智能表征其结构。开发的端到端的AI解析工具，显著提升了谱图鉴定的效率与准确性。如小分子质谱解析方面的DeepMass[78]、Spec2Mol[79]、MIST[53]等；NMR解析方面，SMART 2.0[80]和COLMAR[81]等表现出强大解析能力；离子淌度分析方面，DeepCCS、CCSbase、MetCCS和

AllCC 等模型能够精准预测化合物的 CCS 值[82]，为中药同分异构体的结构鉴定提供支撑；色谱分析方面，Retip[83] 和 RT-Transformer 等模型能够精准预测保留时间。

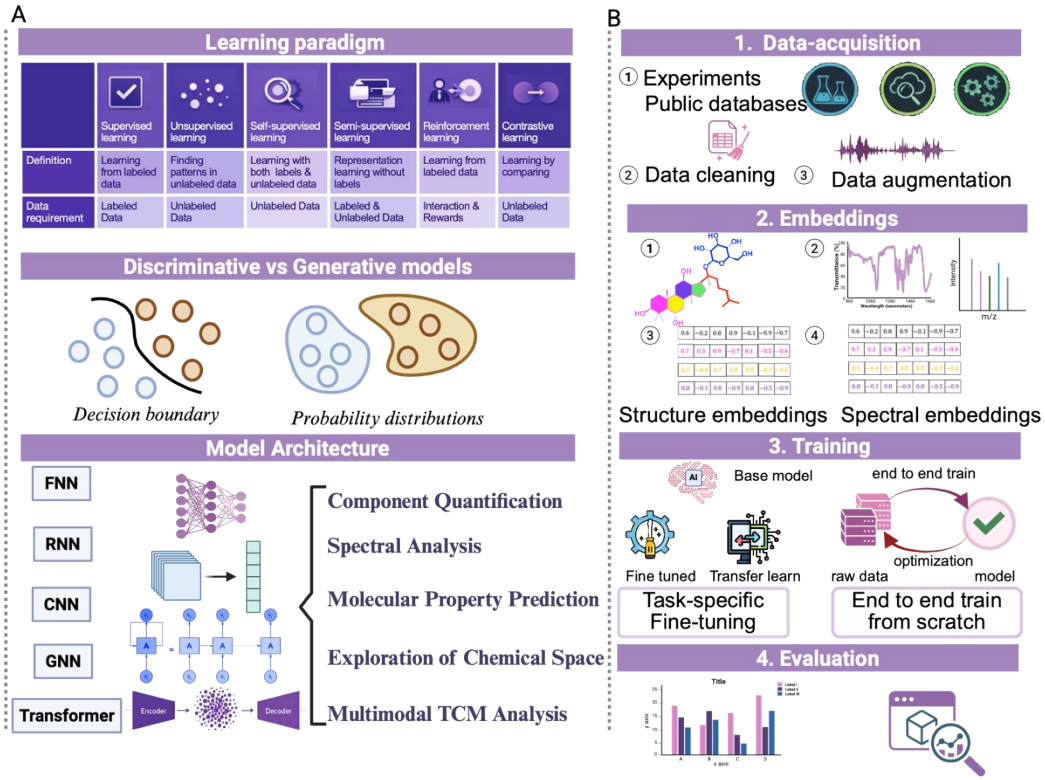

图2　深度学习的范式及主流架构（A）；基于深度学习的中药复杂物质研究的工作流程（B）

Fig. 2　Learning paradigms and mainstream architectures in deep learning（A）；A workflow for deep learning - based studies of complex TCM materials （B）

B consisting of four core stages：（1） data acquisition and preprocessing， including experimental collection， data cleaning， and augmentation；（2） feature embedding， covering representations of structural and spectroscopic information；（3） model training strategies， comparing end-to-end training with task-specific fine-tuning；（4） model evaluation（包括四个核心阶段：（1） 数据获取与预处理，包括实验采集、清洗及增强；（2） 特征嵌入，涵盖结构与光谱信息的表征；（3） 模型训练策略，对比了端到端训练与针对特定任务的微调；（4） 模型评估）

长久以来，中药中的非小分子物质（如多糖、多肽、糖肽和蛋白质）因缺乏标准化谱库且结构多变，成为制约中药药效物质基础研究的核心难题。随着 DeepNovo[84] 等 DL 模型的兴起，可借鉴生物大分子领域成熟的模型，结合迁移学习和强化学习高效表征和预测此类物质的结构和形态。

4.2　反向扩展谱库以快速表征（结构→谱图）

基于 AI 技术从化学结构反向预测谱图，可以有效解决谱库覆盖度低的问题。在光谱图预测方面，一种策略是构建谱图和分子结构的映射关系，进而直接预测分子的谱图；另一种策略是构建人工智能-量子化学模型，如 PaiNN[43]、NequIP[44]、Allegro、DimeNet[40] 和 GemNet[41] 等。这些模型结合 AI，能预测高精度的化学性质，如偶极矩导数、极化率、磁屏蔽张量等，进而利用物理化学关系，计算出理论光谱。该方法在保证光谱预测可靠性的同时，大幅提升了计算效率。目前此类方法主要运用在小分子，在面对复杂大分子时，仍面临挑战。

在质谱图预测方面，蛋白质的谱图预测已趋于成熟，如 DeepMass[78] 和 Prosit[64] 等能够精准预测肽段的多维特征。小分子质谱图预测也在持续发展，如 MetFrag[85]、CFM-ID[63]、OpenMS 3[65] 等已能对小分子的质谱碎裂模式进行有效预测。针对复杂的脂质体系，本团队结合酰基链枚举和人工特征构建了心磷脂专属谱库，实现对低丰度心磷脂的智能解析[86]。在此基础上，本团队进一步整合分子网络技术与谱图嵌入，构建了酰基化脂质谱库，发现了新型乳酰化脂质[87]。在糖类的质谱图预测方面，尽管 DeepGlyco[88] 等工具能够预测糖肽的质谱特征，但由于糖类复杂多样的结构，谱图预测仍存在巨大挑战。

4.3 高阶结构的解析及形成机制探索

现代研究发现，中药的有效成分通过非共价作用力自组装成高阶结构（如自组装体、纳米颗粒、胶束等）发挥药效。通过AI建模可辅助预测中药高阶结构的分子组成、形成机制及结构性质，常用方法包括粗粒化模型、全原子分子动力学、反应力场等，用以预测自主装动态过程[89]。随后，结合几何深度学习，对模拟产生的空间构象进行高纬特征提取。例如，PhysNet模型通过引入物理先验知识，构建消息传递机制，能够捕捉原子间的长程静电力、极化效应及空间非共价作用力，为揭示中药高阶结构的物理本质提供了新工具[90]。

4.4 基因驱动的新物质发现

从基因到化学结构的破译是解析中药复杂成分的关键路径。一方面，“非成簇”弥散分布导致难以从基因组中挖掘新结构，而DeepBGC[91]等深度学习工具能够识别出分散但功能相关的代谢基因；另一方面，随着多种关键代谢酶的编码基因已相继被解析，AI能够通过学习代谢酶的空间构象和底物契合度，鉴定出新型代谢物，如NRPS-Preditor2[92]与NRPStransformer[93]等。鉴于此，中药复杂成分解析的重要方向是利用DeepChem等高性能深度学习框架，构建“基因-代谢酶-结构”预测模型，以基因解析驱动新结构发现。图3展示了通用型AI辅助中药成分智能分析与新物质发现的应用。

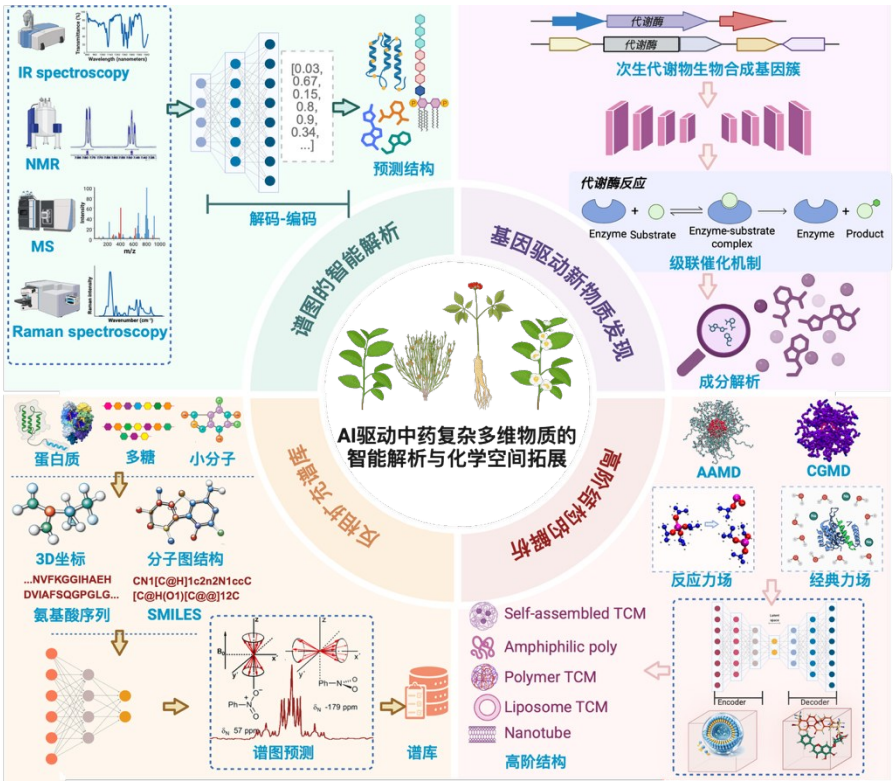

图3 AI驱动中药复杂多维物质成分的智能解析与化学空间拓展

Fig. 3 Artificial intelligence - enabled analysis and chemical space expansion of complex multidimensional constituents in traditional Chinese medicine

为了系统性地评估现有AI工具在中药复杂体系中的适用性，基于文献报道对相关工具在中药复杂物质组解析中的应用进行了总结（表3）。

5 面向液质联用复杂场景的AI辅助解析关键技术

第4部分阐述了通用智能分析框架的构建，但在中药实际研究中，液相色谱-质谱联用技术仍是复杂物质高通量解析的主要手段。该部分将聚焦AI在色谱、质谱分析中的具体场景，分析其应对策略、优势与局限性。

5.1 同分异构体区分策略

中药中存在大量同分异构体，其色谱共洗脱且质谱碎片模式高度一致，传统方法难以有效区分。

结合碰撞截面积和色谱保留时间进行多维特征构建AI模型，为区分同分异构体提供了新途径。然而，由于同分异构体对照品匮乏，基于小样本建立的预测模型缺乏通用性[102]。一方面，可结合迁移学习利用少数对照品数据，在现有的预训练大模型上微调，以提升精准预测能力[103]；另一方面，需要结合GDL提供的空间构象信息，实现多模态数据的深度融合，以弥补实验数据的不足。

表 3　中药复杂物质组智能解析AI工具评估

Table 3　Evaluation of AI tools for intelligent analysis of complex material systems in traditional Chinese medicine

Category	Tool name	Core task	Applicability, advantages & limitations
质谱解析	CFM-ID 4.0[63]	预测质谱谱图；辅助化合物鉴定	在化学逻辑约束下，通过分子图表征与质谱裂解规则建模，具有可解释性。仅支持ESI-MS/MS数据，准确性与计算效率仍有待提升
	SIRIUS[5]	通过质谱谱图预测化合物结构	集合ZODIAC、CSI：FingerID和MSNovelist等模块，实现从分子式确定到全新结构解析的完整分析。但是对训练不足的化合物类别，泛化能力不足
	MS2Query[55]	通过质谱谱图搜索结构类似物	可识别并推测结构类似物及修饰物，适用于已知骨架的衍生扩展；对全新骨架化合物的识别能力有限
	Spec2Mol[79]	通过质谱谱图直接推测结构	不依赖于现有数据库即可输出全新结构，适合未知成分的探索性研究，理念新颖。作为较新的方法，其成熟度和稳定性需大量检验
	MetFrag[85]	基于质谱的结构筛选	整合贝叶斯统计学习评分，对候选结构排序，支持多维度评分；但无法识别新骨架化合物
NMR解析	SMART 2.0[80]	通过NMR谱图预测和分类结构	快速解析谱图，辅助识别结构新颖的天然产物。仅支持HSQC谱图输入，对混合物样品的处理能力有限
基因组挖掘	DeepBGC[91]	从微生物基因组预测合成基因簇	计算效率高，能够发现新颖、非典型的生物合成基因簇。主要针对细菌等原核生物，对植物（中药）基因组的适用性较弱
	antiSMASH[94]	自动化识别、注释基因组中的合成基因簇	基于已知规则的工具，其集成了大量专业的预测模块；且包含针对植物的模块。仅识别经典合成通路，且依赖大量计算资源
分子表征与大模型	Mol2vec[30]	基于1D序列的分子表征学习	方法简单高效，是早期将NLP技术引入化学信息学的经典代表
	ChemBERTa[10]	基于1D序列分子表征学习	基于Transformer的基础模型，具有深层语义提取能力
	DMPNN / Graphormer[95-96]	基于图神经网络架构的表征学习	DMPNN是图神经网络的经典；Graphormer是基于图Transformer的前沿方法，能精准捕捉局部化学环境。但模型训练复杂，对计算资源有要求
底层开发库	DeepChem	化学信息学与机器学习开发库	集成多种分子表示方法，支持多任务学习。需要熟悉Python语言
	RDKit	化学信息学工具包	分子处理库，支持构象生成、指纹计算、子结构搜索等
全流程分析平台	GNPS / FBMN[97-98]	质谱分子网络与全流程分析平台	擅长构建基于特征的分子网络，能发现结构相似的化合物家族，广泛用于天然产物研究
	MZmine/ MS-DIAL[99-100]	LC-MS数据全流程处理平台	开源且功能全面。MZmine参数设置较复杂；MS-DIAL擅长处理脂质组学分析，处理大规模数据时，对计算机内存有较高要求
	Tidymass[101]	全流程工作平台	提供从数据清洗、归一化到统计分析的完整工作流。需要熟悉R语言，主要侧重于数据处理

5.2　噪声干扰的识别与扣除

由于检测困难或功效未明，部分中药成分被视为非目标的“本底信号”，其易混杂于噪音中甚至被误判为噪音。中药无法通过空白背景扣除的方法排除干扰，传统的质量亏损过滤策略亦只能保留中药信号，无法真正识别噪音。因此区分外源噪音（溶剂峰、增塑剂、仪器信号）、随机噪音与中药“本底信号”成为核心难点。而基于AI的对比学习，可利用不同来源的中药构建正样本与负样本的对比矩阵，实现智能识别并扣除偶发信号与外源噪音，同时精准保留药材共有的本底信号[104]。

5.3 低丰度信号的识别与增强

中药成分如鞣质、寡糖与多肽等，因质谱电离效率低、信号丰度弱，且易形成多电荷或复杂加合离子，导致质谱谱图难以解析，甚至无法触发二级质谱采集。面对这种情况：一是将AI整合进数据依赖性采集，实时识别并扣除共有或噪音信号，使低丰度信号达到二级谱图的触发阈值从而获取高质量的二级谱图[105]；二是结合深度学习的超分辨率重建算法，在信号处理层面进一步增强低丰度离子的特征，重构谱图以解析数据[106]。

5.4 复杂离子的解析

中药成分在电喷雾电离过程中易形成加合离子、多电荷离子及源内裂解碎片，显著增加谱图复杂度与假阳性风险。利用 MS-DIAL 等预处理工具已能够有效识别共流出的加合离子簇与同位素峰簇，最新综述也总结了如 IMM 等用于排除冗余节点的方法[107]。在理论预测方面，已有研究尝试对目标结构在电喷雾过程中的电离效率[108]、加合倾向[109]、化学键断裂概率（如 ICEBERG）以及正负离子模式下的统一分析[110]进行建模，以辅助复杂离子解析。尽管该方向正处于快速发展阶段，但目前尚缺乏基于深度图神经网络、覆盖多种加合离子形式的通用模型。

5.5 校正基质效应准确定量

在液质联用的中药定量分析中，离子抑制与基质效应是制约准确定量的瓶颈。IROA TruQuant 通过构建同位素标记内标库并结合数学算法，验证了系统性校正离子抑制的可行性，也为后续基于AI的方法提供了参考[111]。受此启发，有研究采用主动学习策略构建了ESI电离效率预测模型，通过校正质谱响应实现了对天然产物的定量分析[108]。然而，现有模型的泛化能力仍有待提升，构建覆盖常见中药基质类型的大规模训练集成为进一步突破的关键。

6 构建AI智能分析系统存在的挑战及解决方法

6.1 数据共享是基础保障

AI辅助中药分析大模型的训练依赖于大规模高质量数据。因此，数据稀缺与分散问题成为首要挑战：一方面，实验室产生的大量数据未纳入公共数据仓；另一方面，单一实验室的样本多样性有限，数据集价值有限。对此，可借鉴蛋白质组和代谢组学领域的经验，针对中药领域，解决方案包括：（1）借鉴ProteomeXchange和GNPS等平台模式，建立覆盖谱图数据和化学成分数据的中药集中存储仓（区别于现有仅汇总化学成分的数据库）；（2）推出“中药数据论文”发表规范，激励研究者公开并完善数据集元数据（如样本信息和实验参数等），提升数据再利用价值；（3）加强平台质量建设，通过数据质控机制（如谱图质量评估参数）吸引研究者主动上传数据。

6.2 数据标准化是必要前提

各数据库格式规范不统一，导致数据甄别与清洗工作繁重，制约了中药AI模型的构建。借鉴组学领域推出的代谢组学标准计划和蛋白质组学标准计划，可从样本、结构、谱图3个层面构建中药数据标准化体系。

样本元数据方面，在“样本-数据关系”（SDRF）格式[112]的基础上拓展形成SDRF-TCM标准，规范药材产地、炮制方法、提取溶剂及分析参数等特有字段的记录；在中药结构数据方面，引入SMILES、InChIKey等标准化分子描述语言；在谱图数据方面，采用通用谱图标识符为每张谱图赋予唯一的数字标识，支持跨数据库的定位与检索。

6.3 标准化数据需要整合，实现互操性

数据标准化是实现互操作性的必要前提。在语法层面，元数据标准和谱图标识符为跨库整合奠定了基础。在语义层面，还需建立一套领域共通语言，统一规范领域内的词汇表（如岷阳当归和岷县当归）、本体论（如毛蕊花糖苷属于苯乙醇苷类还是酚酸类）和实验方案信息。在技术层面，需要通过开放的交换标准（如API接口等），定义清晰的查询接口，实现机器对机器的数据自动抓取和更新。

6.4 多模态整合解决数据稀缺难题

目前，构建多源数据融合的智能系统是必然趋势。通过整合谱图数据、先验知识、文本数据、组

学数据和成像数据的多模态融合，可实现信息交互，提高解析效率和准确度。（1）谱图与化学先验知识融合。单纯的谱图匹配缺乏化学逻辑约束。本研究团队开发的整合谱图相似性与碎片注释的评分策略，融合了化学先验知识与谱图数据，通过对皂苷复杂组分进行量化置信度评估，有效提升了中药复杂成分解析的准确度[113]。（2）谱图与文本融合。如结合气相离子化学理论与检索增强生成技术，构建光谱-语言融合大模型，实现分子结构自动表征[114]及检测条件优化等任务的自然语言交互[115]；（3）多组学数据融合，将植物基因组、代谢组和酶功能等整合至统一的化学信息学框架中，构建从代谢酶到化学物质的发现路径[116]。

参考文献：

[1] Huber F, Ridder L, Verhoeven S, Spaaks J H, Diblen F, Rogers S, van der Hooft J J J. PLoS Comput. *Biol.*, **2021**, 17 (2): e1008724.
[2] Martin M R, Hassoun S. *Anal. Chem.*, **2025**, 97 (38): 20734-20742.
[3] Mokaya M, Imrie F, van Hoorn W P, Kalisz A, Bradley A R, Deane C M. *Nat. Mach. Intell.*, **2023**, 5 (4): 386-394.
[4] Zhang Q, Qu J F, Y, Zhou H Y. J. Instrum. Anal. (张琦, 区锦锋, 周华英.分析测试学报), **2024**, 47 (3): 29-33
[5] Böcker S, Letzel M C, Lipták Z, Pervukhin A. *Bioinformatics*, **2009**, 25 (2): 218-224.
[6] Sohn K, Berthelot D, Li C-L, Zhang Z, Carlini N, Cubuk E D, Kurakin A, Zhang H, Raffel C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence, 2020. [2025-12-10]. https://doi.org/10.48550/arXiv.2001.07685.
[7] Bushuiev R, Bushuiev A, Samusevich R, Brungs C, Sivic J, Pluskal T. *Nat. Biotechnol.*, **2025**: doi: 10.1038/s41587-41025-02663-41583.
[8] Wang Z Y, Mi J C, Lu S, He J Y. [2026-1-3]. https://doi.org/10.48550/arXiv.2311.16666.
[9] Loeffler H H, He J, Tibo A, Janet J P, Voronov A, Mervin L H, Engkvist O. J. *Cheminform.*, **2024**, 16 (1): 20.
[10] Chithrananda S, Grand G, Ramsundar B. [2026-1-13]. https://arxiv.org/pdf/2010.09885.
[11] Xie M Y, Shan S N S, Li H H, Cheng W X. J. *Anhui Univ. Chin. Med.* (解玫莹, 单圣男, 李欢欢, 程旺兴.安徽中医药大学学报), **2021**, 40 (6): 86-91.
[12] Horgan C C, Jensen M, Nagelkerke A, St-Pierre J P, Vercauteren T, Stevens M M, Bergholt M S. *Anal. Chem.*, **2021**, 93 (48): 15850-15860.
[13] Intan A, Zetta D, Jarukamjorn K, Srisongkram T. *ACS Omega*, **2025**, 10 (38): 43616-43631.
[14] Gananathan K, Manjula D, Sugumaran V. BMC. *Bioinform.*, **2025**, 26 (1): 185.
[15] Therrien F, Sargent E H, Voznyy O. *Nat. Commun.*, **2025**, 16 (1): 4301.
[16] Masood M A, Kaski S, Cui T. *J. Cheminform.*, **2025**, 17 (1): 58.
[17] Singhal K, Azizi S, Tu T, Mahdavi S S, Wei J, Chung H W, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera Y A B, Webster D, Corrado G S, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. *Nature*, **2023**, 620 (7972): 172-180.
[18] Zhou R R, Wang Y K, Zhen L P, Shen B B, Long H P, Huang L Q. *Foods*, **2025**, 14 (8): 1340.
[19] Gao M M, Huang J Y, Zhai Y N, Jiang L F, Lu G S, Hu X X, Ge X Q. *Chin J. Pharm. Anal.* (高美美, 黄建猷, 翟雅南, 蒋凌风, 陆国寿, 胡筱希, 雪晴, 李冬梅.药物分析杂志), **2024**, 44 (5): 882-892.
[20] Stravs M A, Dührkop K, Böcker S, Zamboni N. *Nat. Methods*, **2022**, 19 (7): 865-870.
[21] Yu N Y, Ma Z, Shao Q, Li L H, Wang X B, Pan B C, Yu H X, Wei S. *Nat. Mach. Intell.*, **2025**, 7 (11): 1879-1887.
[22] Lan T, Su S Q, Ping P Y, Hutvagner G, Liu T, Pan Y, Li J Y. *Nat. Mach. Intell.*, **2024**, 6 (3): 315-325.
[23] Inukai T, Yamato A, Akiyama M, Sakakibara Y. *Commun. Chem.*, **2025**, 8 (1): 228.
[24] Jiang Y Y, Zhang G, You J, Zhang H L, Yao R, Xie H Z, Zhang L Y, Xia Z Y, Dai M Z, Wu Y J, Li L L, Yang S Y. *Nat. Mach. Intell.*, **2024**, 6 (3): 326-337.
[25] Watson J L, Juergens D, Bennett N R, Trippe B L, Yim J, Eisenach H E, Ahern W, Borst A J, Ragotte R J, Milles L F, Wicky B I M, Hanikel N, Pellock S J, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres S V, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola T S, DiMaio F, Baek M, Baker D. *Nature*, **2023**, 620 (7976): 1089-1100.
[26] Lin X, Xia Y, Li Y, Huang Y P, Liu S, Zhang J, Gao Y Q. *Nat. Commun.*, **2025**, 16 (1): 6043.
[27] Zong G, Gao J, Qi Y, Zhao H, Feng W, Hu B, Ma J, Du L, Han J. *Anal. Chem.*, **2025**, 97 (49): 27325-27336.
[28] Zhu J J, Miao S Q, Ying R, Li P. *Nat. Mach. Intell.*, **2025**, 7 (3): 471-483.

[29] Castro Nascimento C M, Pimentel A S. *J. Chem. Inf. Model*, **2023**, 63 (6): 1649-1655.
[30] Jaeger S, Fulle S, Turk S. *J. Chem. Inf. Model*, **2018**, 58 (1): 27-35.
[31] He X H, Zhao L F, Tian Y P, Li R, Chu Q Y, Gu Z Y, Zheng M Y, Wang Y S, Li S N, Jiang H L, Jiang Y, Wen L Q, Wang D Y, Cheng X. *Nat. Commun.*, **2024**, 15 (1): 5163.
[32] Abtheen E A, Singh A, Sriram S, Chen C, Neelamegham S, Gunawan R. [2015-12-21]. https://doi.org/10.1101/2025.07.02.662857.
[33] Liu X, Ye K, van Vlijmen H W T, AP I J, van Westen G J P. J. *Cheminform.*, **2023**, 15 (1): 24.
[34] Méndez-Lucio O, Nicolaou C A, Earnshaw B. *Nat. Commun.*, **2024**, 15 (1): 9431.
[35] Li H, Zhang R, Min Y, Ma D, Zhao D, Zeng J Y. *Nat. Commun.*, **2023**, 14 (1): 7568.
[36] Young A, Wang B, Röst H. [2025-12-24]. https://doi.org/10.48550/arXiv.2111.04824.
[37] Alkuhlani A, Gad W, Roushdy M, Salem A-B M. *IEEE Access*, **2023**, 11: 51838-51847.
[38] Burkholz R, Quackenbush J, Bojar D. *Cell Rep.*, **2021**, 35 (11): 109251.
[39] Joeres R, Bojar D. [2025-12-21].https://doi.org/10.48550/arXiv.2409.13467.
[40] Gasteiger J, Groß J, Günnemann S. [2025-12-24].https://doi.org/10.48550/arXiv.2003.03123.
[41] Gasteiger J, Becker F, Günnemann S. [2025-12-27]. https://doi.org/10.48550/arXiv.2106.08903.
[42] Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K, Riley P. [2025-12-30]. https://doi.org/10.48550/arXiv.1802.08219.
[43] Schütt K T, Unke O T, Gastegger M. [2025-12-30]. https://doi.org/10.48550/arXiv.2102.03150.
[44] Batzner S, Musaelian A, Sun L, Geiger M, Mailoa J P, Kornbluth M, Molinari N, Smidt T E, Kozinsky B. *Nat. Commun.*, **2022**, 13: 2453.
[45] Réau M, Renaud N, Xue L C, Bonvin A M J J. *Bioinformatics*, **2022**, 39 (1): btac759.
[46] Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein M M, Correia B E. *Nat. Methods*, **2020**, 17 (2): 184-192.
[47] Roche R, Moussad B, Shuvo M H, Bhattacharya D. PLoS Comput. *Biol.*, **2023**, 19 (8): e1011435.
[48] Xu C S, Deng X M, Lu Y, Yu P Y. *Digit. Discov.*, **2024**, 4 (1): 161-171.
[49] Morehead A, Cheng J. *Commun. Chem.*, **2024**, 7 (1): 150.
[50] Sheshanarayana R, You F. *Digit. Discov.*, **2025**, 4: 2298-2335.
[51] Kirchmeyer M, Pinheiro P O, Willett E, Martinkus K, Kleinhenz J, Makowski E K, Watkins A M, Gligorijevic V, Bonneau R, Saremi S. [2026-1-15]. https://doi.org/10.48550/arXiv.2511.15906.
[52] Kelly T, Xia S, Lu J, Zhang Y. *J. Chem. Inf. Model*, **2025**, 65 (8): 3990-3998.
[53] Goldman S, Wohlwend J, Stražar M, Haroush G, Xavier R J, Coley C W. *Nat. Mach. Intell.*, **2023**, 5 (9): 965-979.
[54] Huber F, van der Burg S, van der Hooft J J J, Ridder L. J. *Cheminform.*, **2021**, 13 (1): 84.
[55] de Jonge N F, Louwen J J R, Chekmeneva E, Camuzeaux S, Vermeir F J, Jansen R S, Huber F, van der Hooft J J J. *Nat. Commun.*, **2023**, 14 (1): 1752.
[56] Altenburg T, Giese S, Wang S, Muth T, Renard B. *Nat. Mach. Intell.*, **2022**, 4 (4): 378-388
[57] Xie T, Zhang H L, Yang Q, Sun J Y, Wang Y, Long J, Zhang Z M, Lu H M. *Anal. Chem.*, **2025**, 97 (25): 13350-13360.
[58] Xiong P, Xu H, Zheng H. *Anal. Chem.*, **2025**, 97 (37): 20137-20146.
[59] Abdelmoula W M, G-CLopez B, Randall E C, Kapur T, Sarkaria J N, White F M, Agar J N, Wells W M, Agar N Y R. *Nat. Commun.*, **2021**, 12 (1): 5544.
[60] Abdelmoula W M, Stopka S A, Randall E C, Regan M, Agar J N, Sarkaria J N, Wells W M, Kapur T, Agar N Y R. *Bioinformatics*, **2022**, 38 (7): 2015-2021.
[61] Guo L, Xie C Y, Miao R, Xu J J, Xu X N, Fang J C, Wang X X, Liu W P, Liao X W, Wang J N, Dong J Y, Cai Z W. *Anal. Chem.*, **2024**, 96 (9): 3829-3836.
[62] Borodinov N, Lorenz M, King S T, Ievlev A V, Ovchinnikova O S. *Npj Comput. Mater.*, **2020**, 6 (1): 83.
[63] Wang F, Liigand J, Tian S, Arndt D, Greiner R, Wishart D S. *Anal. Chem.*, **2021**, 93 (34): 11692-11700.
[64] Gabriel W, González R M, Laposchan S, Riedel E, Dündar G, Poppenberger B, Wilhelm M, Lee C-Y. Mol. *Cell Proteomics*, **2025**, 24 (3): 100924.
[65] Pfeuffer J, Bielow C, Wein S, Jeong K, Netz E, Walter A, Alka O, Nilse L, Colaianni P D, McCloskey D, Kim J, Rosenberger G, Bichmann L, Walzer M, Veit J, Boudaud B, Bernt M, Patikas N, Pilz M, Startek M P, Kutuzova S, Heumos L, Charkow J, Sing J C, Feroz A, Siraj A, Weisser H, Dijkstra T M H, Perez-Riverol Y, Röst H, Kohlbacher O, Sachsenberg T. *Nat. Methods*, **2024**, 21 (3): 365-367.
[66] Gloaguen Y, Kirwan J A, Beule D. *Anal. Chem.*, **2022**, 94 (12): 4930-4937.
[67] Li H, Long D, Yuan L, Wang Y, Tian Y H, Wang X C, Mo F. *Nat. Comput. Sci.*, **2025**, 5 (3): 234-244.
[68] Li D W, Hansen A L, Yuan C, Bruschweiler-Li L, Brüschweiler R. *Nat. Commun.*, **2021**, 12 (1): 5229.
[69] Wu W, Leonardis A, Jiao J, Jiang J, Chen L. *J. Phys. Chem. A*, **2025**, 129 (8): 2077-2085.
[70] Stienstra C M K, Hebert L, Thomas P, Haack A, Guo J, Hopkins W S. *J. Chem. Inf. Model*, **2024**, 64 (12):

4613-4629.
[71] Luo Y, Chen W H, Su Z H, Shi X Q, Luo J, Qu X B, Chen Z, Lin Y Q. *Nat. Commun.*, **2025**, 16: 2342.
[72] Tian Z J, Dai Y, Hu F, Shen Z H, Xu H L, Zhang H W, Xu J H, Hu Y T, Diao Y Y, Li H L. *J. Chem. Inf. Model*, **2024**, 64 (14): 5624-5633.
[73] Tan X. *J. Cheminform.*, **2025**, 17 (1): 103.
[74] Yang Y J, Zheng J, Guo P, Gao Q, Guo Y W, Chen Z, Liu C C, Wu T Q, Ouyang Z L, Chen H, Kang Y. *Front. Artif. Intell.*, **2025**, 8: 1466643.
[75] Klukowski P, Riek R, Güntert P. *Nat. Commun.*, **2022**, 13 (1): 6151.
[76] Kazemzadeh M, Martinez-Calderon M, Xu W, Chamley L W, Hisey C L, Broderick N G R. *Anal. Chem.*, **2022**, 94 (37): 12907-12918.
[77] Georgiev D, Fernández-Galiana Á, Vilms Pedersen S, Papadopoulos G, Xie R, Stevens M M, Barahona M. *Proc. Natl. Acad. Sci.*, **2024**, 121 (45): e2407439121.
[78] Ji H C, Du R, Dai Q L, Su M F, Lyu Y Q, Peng Y C, Yan J B. *bioRxiv*, 2024: 596727.
[79] Litsa E E, Chenthamarakshan V, Das P, Kavraki L E. *Commun. Chem*, **2023**, 6 (1): 132.
[80] Reher R, Kim H W, Zhang C, Mao H H, Wang M, Nothias L F, Caraballo-Rodriguez A M, Glukhov E, Teke B, Leao T, Alexander K L, Duggan B M, Van Everbroeck E L, Dorrestein P C, Cottrell G W, Gerwick W H. *J. Am. Chem. Soc.*, **2020**, 142 (9): 4114-4120.
[81] Li D W, Cabrera Allpas R, Choo M, Bruschweiler-Li L, Hansen A L, Brüschweiler R. Anal .*Chem.*, **2024**, 96 (43): 17174-17183.
[82] Li X H, Wang H D, Jiang M T, Ding M X, Xu X Y, Xu B, Zou Y D, Yu Y T, Yang W Z. *Molecules*, **2023**, 28 (10): 4050.
[83] Kuhn M. *J. Stat. Soft*, **2008**, 28 (5): 26.
[84] Tran N H, Zhang X, Xin L, Shan B, Li M. *Proc. Natl. Acad. Sci. USA*, **2017**, 114 (31): 8247-8252.
[85] Ruttkies C, Neumann S, Posch S. *BMC Bioinform.*, **2019**, 20 (1): 376.
[86] He Y B, Yuan B, Lu Y, Zhao X, Shen C, Ji J J, Lin L, Xu J, Xie T, Shan J J. *Anal. Chim. Acta*, **2021**, 1180: 338879.
[87] Yuan B, Li X, Xu S, Sun H, Shen C, Ji J J, Lin L L, Xu W C, Shan J J, Tong W, Xie T. *Anal. Chem.*, **2023**, 95 (22): 8443-8451.
[88] Yang Y, Fang Q. *Nat. Commun.*, **2024**, 15 (1): 2448.
[89] Ge Y, Zhu Q, Wang X, Ma J. *Ind. Chem. Mater.*, **2025**, 3 (4): 383-411.
[90] Unke O T, Meuwly M. *J. Chem. Theory Comput.*, **2019**, 15 (6): 3678-3693.
[91] Hannigan G D, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, Durcak J, Wurst M, Kotowski J, Chang D, Wang R, Piizzi G, Temesi G, Hazuda D J, Woelk C H, Bitton D A. Nucleic Acids. *Res.*, **2019**, 47 (18): e110.
[92] Röttig M, Medema M H, Blin K, Weber T, Rausch C, Kohlbacher O. Nucleic. Acids. *Res.*, **2011**, 39: W362-W367.
[93] Zhang Z, Zhou Y, Xie S, Liu R-Z, Huang Z, Saravana Kumar P, Feng G, Yuan F, Zhang L. *J. Am. Chem. Soc.*, **2025**, 147 (36): 32662-32670.
[94] Blin K, Shaw S, Vader L, Szenei J, Reitz Z L, Augustijn H E, Cediel-Becerra J D D, de Crécy-Lagard V, Koetsier R A, Williams S E, Cruz-Morales P, Wongwas S, Segurado Luchsinger A E, Biermann F, Korenskaia A, Zdouc M M, Meijer D, Terlouw B R, van der Hooft J J J, Ziemert N, Helfrich E J N, Masschelein J, Corre C, Chevrette M G, van Wezel G P, Medema M H, Weber T. *Nucleic. Acids Res.*, **2025**, 53 (W1): W32-W38.
[95] Wang Z, Oh S-K, Fu Z, Roh S-B, Pedrycz W. *Eng. Appl. Artif. Intell.*, **2025**, 157: 111164.
[96] Stienstra C M K, Hebert L, Thomas P, Haack A, Guo J, Hopkins W S. *J. Chem. Inf. Model*, **2024**, 64 (12): 4613-4629.
[97] Wang M, Carver J J, Phelan V V, Sanchez L M, Garg N, Peng Y, Nguyen D D, Watrous J, Kapono C A, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik A V, Meehan M J, Liu W T, Crüsemann M, Boudreau P D, Esquenazi E, Sandoval-Calderón M, Kersten R D, Pace L A, Quinn R A, Duncan K R, Hsu C C, Floros D J, Gavilan R G, Kleigrewe K, Northen T, Dutton R J, Parrot D, Carlson E E, Aigle B, Michelsen C F, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy B T, Gerwick L, Liaw C C, Yang Y L, Humpf H U, Maansson M, Keyzers R A, Sims A C, Johnson A R, Sidebottom A M, Sedio B E, Klitgaard A, Larson C B, P C A B, Torres-Mendoza D, Gonzalez D J, Silva D B, Marques L M, Demarque D P, Pociute E, O'Neill E C, Briand E, Helfrich E J N, Granatosky E A, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush J J, Zeng Y, Vorholt J A, Kurita K L, Charusanti P, McPhail K L, Nielsen K F, Vuong L, Elfeki M, Traxler M F, Engene N, Koyama N, Vining O B, Baric R, Silva R R, Mascuch S J, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams P G, Dai J, Neupane R, Gurr J, Rodríguez A M C, Lamsa A, Zhang C, Dorrestein K, Duggan B M, Almaliti J, Allard P M, Phapale P, Nothias L F, Alexandrov T, Litaudon M, Wolfender J L, Kyle J E, Metz T O, Peryea T, Nguyen D T, VanLeer D, Shinn P, Jad-

hav A, Müller R, Waters K M, Shi W, Liu X, Zhang L, Knight R, Jensen P R, Palsson B O, Pogliano K, Linington R G, Gutiérrez M, Lopes N P, Gerwick W H, Moore B S, Dorrestein P C, Bandeira N. *Nat. Biotechnol.*, **2016**, 34(8): 828-837.

[98] Ding Y X, He H B, Wang Y, Yuan C Y, Zheng X X, Cao Z, Cai Y P, Xing Y C. *Anal. Bioanal. Chem.*, **2025**, 417(28): 6489-6500.

[99] Heuckeroth S, Damiani T, Smirnov A, Mokshyna O, Brungs C, Korf A, Smith J D, Stincone P, Dreolin N, Nothias L-F, Hyötyläinen T, Orešič M, Karst U, Dorrestein P C, Petras D, Du X, van der Hooft J J J, Schmid R, Pluskal T. Nat. *Protocols*, **2024**, 19(9): 2597-2641.

[100] Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, Kanazawa M, VanderGheynst J, Fiehn O, Arita M. *Nat. Methods*, **2015**, 12(6): 523-526.

[101] Wang X, Liu Y J, Jiang C, Huang Z N, Yan H, Wong S H, Johnson C H, Zhang J X, Ge Y F, Zhang F F, Zhang J L, Lai R F, Gao P, Zhang X B, Shen X T. *Nat. Commun.*, **2026**, 17(1): 1755.

[102] Chen M, Hao Y, Huang X, Wu P, Sun J, Zhang B, Chen S. *Nat. Commun.*, **2026**, 17(1): 2554.

[103] Wang B, Pan B, Zhang T, Liu Q, Li S. *Adv. Sci.*, **2025**, 12(46): e09456.

[104] Yang Y, Li M, Zhang X, Qin Y, Cai C, Liu Z, Zhai D, Li P, Shang J. Spectrochim. *Acta A*, **2026**, 349: 127289.

[105] Wallmann G, Skowronek P, Brennsteiner V, Lebedev M, Thielert M, Steigerwald S, Kotb M, Despard O, Heymann T, Zhou X-X, Strauss M T, Ammar C, Willems S, Schwörer M, Zeng W-F, Mann M. *Nat. Biotechnol.*, **2025**, DOI: 10.1038/s41587-025-02791-w.

[106] Wang J Y, Li B Q, Zhai Y B, Liu L Y, Jiang T, Xu W. *Anal. Chem.*, **2025**, 97(12): 6489-6496.

[107] El Abiead Y, Mohanty I, Xing S, Rutz A, Charron-Lamoureux V, Damiani T, Lu W, Patti G J, Zamboni N, Yanes O, Dorrestein P C. *JACS Au*, **2025**, 5(12): 5828-5850.

[108] Wang W-C, Amini N, Huber C, Kull M, Kruve A. *Anal. Chem.*, **2025**, 97(25): 13131-13139.

[109] Costalunga R, Tshepelevitsh S, Sepman H, Kull M, Kruve A. *Anal. Chim. Acta*, **2022**, 1204: 339402.

[110] de Jonge N F, Chekmeneva E, Schmid R, Joas D, Truong L-J, van der Hooft J J J, Huber F. *Nat. Commun.*, **2026**, 17(1): 2483.

[111] Mahmud I, Wei B, Veillon L, Tan L, Martinez S, Tran B, Raskind A, de Jong F, Liu Y, Ding J, Xiong Y, W-kChan, Akbani R, Weinstein J N, Beecher C, Lorenzi P L. *Nat. Commun.*, **2025**, 16(1): 1347.

[112] Dai C, Füllgrabe A, Pfeuffer J, Solovyeva E M, Deng J, Moreno P, Kamatchinathan S, Kundu D J, George N, Fexova S, Grüning B, Föll M C, Griss J, Vaudel M, Audain E, Locard-Paulet M, Turewicz M, Eisenacher M, Uszkoreit J, Van Den Bossche T, Schwämmle V, Webel H, Schulze S, Bouyssié D, Jayaram S, Duggineni V K, Samaras P, Wilhelm M, Choi M, Wang M, Kohlbacher O, Brazma A, Papatheodorou I, Bandeira N, Deutsch E W, Vizcaíno J A, Bai M, Sachsenberg T, Levitsky L I, Perez-Riverol Y. *Nat. Commun.*, **2021**, 12(1): 5854.

[113] Xie T, Shan J J, Jiang J, Zhao X, He Y, Tong W J. *J. Pharm. Biomed. Anal.*, **2021**, 204: 114291.

[114] Zhao L, Zhou M, Jiang J. *J. Phys. Chem. Lett.*, **2025**, 16(18): 4382-4391.

[115] Lapin J, Iuzhaninov M, Hölzlwimmer A J, Wilhelm M.**2025**. https://doi.org/10.1101/2025.11.19.689259.

[116] Lee B T, Kwon J Y, Weber T, Kim H U. Nat. *Prod. Rep*, **2026**. https://doi.org/10.1039/D1035NP00059A.

(责任编辑:盛文彦)